9-30-2015

Whole-genome sequencing of KSHV from Zambian Kaposi's sarcoma biopsies reveals unique viral diversity

Landon N. Olp
University of Nebraska-Lincoln

Adrien Jeanniard
University of Nebraska-Lincoln

Clemence Marimo
University of Zambia School of Medicine, Lusaka, Zambia

John T. West
University of Nebraska-Lincoln, jwest2@unl.edu

Charles Wood
University of Nebraska-Lincoln, cwood1@unl.edu

JVI Accepted Manuscript Posted Online 30 September 2015
J. Virol. doi:10.1128/JVI.01712-15

Whole-genome sequencing of KSHV from Zambian Kaposi's sarcoma biopsies reveals unique viral diversity.

Landon N. Olp[1], Adrien Jeanniard[1], Clemence Marimo[2], John T. West[1], and Charles Wood[1]#

[1]Nebraska Center for Virology and School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, NE, USA; [2]Department of Pathology and Microbiology, University of Zambia School of Medicine, Lusaka, Zambia

Running title: Whole-genome sequencing of Zambian KSHV

#Address correspondence to Charles Wood, cwood1@unl.edu.

Abstract word count: 247

Text word count: 4,315

ABSTRACT

Kaposi's sarcoma-associated herpesvirus (KSHV) is the etiological agent for Kaposi's sarcoma (KS). Both KSHV and KS are endemic in sub-Saharan Africa where approximately 84% of global KS cases occur. Nevertheless, whole-genome sequencing of KSHV has only been completed using isolates from Western countries—where KS is not endemic. The lack of whole-genome KSHV sequence data from the most clinically important geographical region, sub-Saharan Africa, represents an important gap as it remains unclear whether genomic diversity has a role on KSHV pathogenesis. We hypothesized that distinct KSHV genotypes might be present in sub-Saharan Africa compared to Western countries. Using a KSHV-targeted enrichment protocol followed by Illumina deep-sequencing, we generated and analyzed sixteen unique Zambian, KS-derived, KSHV genomes. We enriched KSHV DNA over cellular DNA 1,851 to 18,235-fold. Enrichment provided coverage levels up to 24,740-fold; therefore, supporting highly confident polymorphism analysis. Multiple alignment of the sixteen newly sequenced KSHV genomes showed low level variability across the entire central conserved region. This variability resulted in distinct phylogenetic clustering between Zambian KSHV genomic sequences and those derived from Western countries. Importantly, the phylogenetic segregation of Zambian from Western sequences occurred irrespective of inclusion of the highly variable genes K1 and K15. We also show that four genes within the more conserved region of the KSHV genome contained polymorphisms that partially, but not fully, contributed to the unique Zambian KSHV whole-genome phylogenetic structure. Taken together, our data suggest that the whole KSHV genome should be taken into consideration for accurate viral characterization.

IMPORTANCE

Our results represent the largest number of KSHV whole-genomic sequences published to date and the first time that multiple genomes have been sequenced from sub-Saharan Africa, a geographic area where KS is highly endemic. Based on our new sequence data, it is apparent that whole-genome KSHV diversity is greater than previously appreciated and differential phylogenetic clustering exists between viral genomes of Zambia and Western countries. Furthermore, individual genes may be insufficient for KSHV genetic characterization. Continued investigation of the KSHV genetic landscape is necessary in order to effectively understand the role of viral evolution and sequence diversity on KSHV gene functions and pathogenesis.

INTRODUCTION

Kaposi's sarcoma-associated herpesvirus (KSHV), or human herpesvirus-8 (HHV-8), is the etiologic agent for all forms of Kaposi's sarcoma (KS) (1). KS manifests as an endothelial tumor primarily on the skin but can also involve mucosal membranes and visceral organs. Among the HIV-uninfected population KS is rare worldwide; however, HIV infection and immunosuppression greatly increase the risk of developing KS (2). In sub-Saharan Africa, HIV is epidemic and KSHV is endemic. Accordingly, KS is one of the most common cancers in sub-Saharan Africa and this region accounts for 84% of global KS cases (3). Two other HIV-associated lymphoproliferative malignancies (primary effusion lymphoma [PEL] and multicentric Castleman's disease [MCD]), as well as the KSHV inflammatory cytokine syndrome (KICS) are also associated with KSHV infection (4-6). However, the role of KSHV genetic variation on pathogenesis and disease presentation is unknown. Therefore, as a first step, it is necessary to analyze KSHV genetic variation in sub-Saharan Africa at the whole-genome level.

KSHV is a human gamma-herpesvirus with a largely conserved double-stranded DNA genome of approximately 140 kilobases. However, the extreme 5' and 3' termini are disproportionately variable compared to the central region of the KSHV genome and both have been used to categorize KSHV into different genotypes (7, 8) The 5' end encodes the K1 gene and can be separated into five distinct genotypes (A, B, C, D, E), differing by up to 30% at the amino acid level. At the nucleotide level, 85% of polymorphisms within K1 are nonsynonymous, suggesting that strong selective pressure acts on the gene (7). The 3' terminus of the KSHV genome encodes the K15 gene. Sequence analysis of K15 supports additional categorization of KSHV sequences

Journal of Virology

into P, M, or N alleles, with up to 70% inter-allele divergence at the amino acid level (8, 9). In addition, nine discrete loci (≈5.6% of the genome) within the central, more conserved, region of the KSHV genome also contain polymorphisms, albeit at a much lower rate. Together, twelve KSHV genotypes have been proposed based on these 11 discrete loci (9). However, the remaining KSHV genes, representing more than 90% of the genome, have not been used to further characterize KSHV genetic structure and diversity due to a lack of high coverage, whole-genome, viral sequences.

Presently, only six KSHV whole-genome sequences are available. The first complete, and most extensively annotated genome, GK18, was generated from a classic KS lesion from a Greek patient (AF148805.2) (10). The nearly complete, 'KS' genome was sequenced using shotgun sequencing of fragments obtained after Sau3A digestion of DNA from AIDS-associated KS biopsies (U93872.2) (11). Additionally, three genomic KSHV sequences were generated from KSHV-infected PEL cell lines, BC-1 (U75698.1), JSC-1 (GQ994935.1), and BCBL-1 (HQ404500.1) (12-14). The sixth and most recently sequenced KSHV genome, DG-1 (JQ619843.1), was the first to be completed using Illumina next-generation sequencing technology and the first obtained from virus in patient plasma (15). Despite these significant efforts, all current genomic KSHV sequences were generated from samples obtained in Western countries where KSHV is not endemic. The lack of whole-genome KSHV sequence data from sub-Saharan Africa—the geographical region most relevant to KSHV infection—represents an important gap in genetic characterization for this pathogen as it remains unclear whether a correlation exists between whole-genome sequence diversity and KSHV pathogenesis. A recent study of Epstein Barr Virus (EBV) whole genomes revealed significant levels of sequence

diversity in isolates from nasopharyngeal carcinoma (NPC) clinical samples in a region with high NPC prevalence (16). This further suggests that a thorough characterization of KSHV whole-genomes needs to be conducted—including isolates from sub-Saharan Africa— as a first step to investigate possible relationships between genomic diversity and pathogenesis.

In the current study, we sought to test the hypothesis that distinct whole-genome KSHV variants are present in sub-Saharan Africa compared to Western countries. We also investigated whether diversity within the central region genes may contribute to viral characterization. To this end, we generated and analyzed KSHV whole-genome sequences derived from KS skin lesions of sixteen different Zambian patients. Using a biotinylated RNA-library as bait, KSHV sequences were preferentially enriched over human genomic DNA present in tumor samples and the KSHV-enriched DNA was sequenced using Illumina deep-sequencing technology. Polymorphism and phylogenetic analyses were then performed to measure KSHV genome-wide diversity. Our results represent the largest number of KSHV whole-genomic sequences published to date and the first time that multiple genomes have been sequenced from sub-Saharan Africa, a geographic area where KS is highly endemic.

METHODS

KS sample collection. KS biopsies were obtained from patients upon disease presentation at the skin clinic of the University of Zambia, University Teaching Hospital. The biopsies were collected as part of KS diagnosis and residual tissue samples were used for the current study. All patients provided written, informed consent to participate in the study. Collection of biopsies was

approved by the Institutional Review Board of the University of Nebraska and the University of Zambia Biomedical Research Ethics Committee.

Sample DNA preparation. Total DNA was extracted from frozen KS tumor biopsies using the Gentra Puregene Tissue Kit according to manufacturer's protocol (Qiagen). Purified DNA samples were analyzed via Qubit broad-range dsDNA kit, Nanodrop spectrometer, and agarose gel electrophoresis to measure concentration, protein contamination, and level of degradation. All samples were of high quality and usable for downstream applications.

KSHV viral load. The number of KSHV genomes in each KS biopsy sample was quantified before enrichment using the Bio-Rad QX100 droplet digital PCR (ddPCR) system. Human β-globin and KSHV ORF26 were amplified using primers and probes described previously (17). Each 20μL ddPCR reaction contained 1X ddPCR Supermix (Bio-Rad), 900nM forward and reverse primer, 250nM Taqman probe, and 6 or 60ng of genomic DNA. Droplet generation, amplification, and reading were carried out according to the manufacturer's protocol. Amplification conditions for β-globin were as follows: 95°C for 10 min, 40 cycles of 94°C for 30 sec and 60°C for 60 sec, and 98°C for 10 min. KSHV ORF26 amplification conditions were similar with the exception of a 55°C elongation temperature. All samples were run in triplicate and the mean KSHV copy number per cellular equivalent was calculated.

Library preparation, Target Enrichment, and Illumina sequencing. Sample library preparation and target enrichment were performed using the SureSelect^{XT} Target Enrichment System (Agilent) according to the manufacturer's Illumina paired-end sequencing library

protocol. Briefly, 120bp overlapping RNA baits were custom designed at 5x coverage in conjunction with Agilent SureDesign based on the KSHV GK18 sequence (AF148805.2). Baits with high homology to human DNA were excluded from the RNA-library. For each sample, 3µg of purified KS tissue biopsy DNA was sheared and Illumina specific adapters were added. The DNA-libraries were then hybridized for KSHV-specific enrichment, index tagged, and pooled at equimolar amounts for sequencing. Next-generation sequencing was performed using an Illumina HiSeq with 100bp paired-end reads on two separate runs at the University of Nebraska DNA Sequencing Core.

Guided assembly of KSHV genomes. Output reads from the Illumina HiSeq 2500 were filtered using Trimmomatic (18). Reads were trimmed on both on both 5' and 3' extremities using a quality (Q) threshold keeping only bases with Q>=30. All reads shorter than 101bp long were filtered out, thus retaining only full-size reads of high quality. The dataset was controlled for quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), both before and after the filtering steps.

High quality paired-end reads were aligned to the KSHV genome (GK18, accession: AF148805.2) using Bowtie2 version 2.2.1 (19). First, reads were assembled one sample at a time using the Columbus extension of Velvet 1.2.10 (20) and the corresponding initial alignment. A k-mer size of 91 was found to produce the best results after multiple trials of different values. Both average and minimum k-mer coverage were determined on a sample-by-sample basis by first performing a 'blank' assembly with the 'exp_cov' and 'cov_cutoff' parameters set on 'auto' and then looking at the k-mer coverage of the largest contigs produced matching the reference

sequence. Afterwards, the 'cov_cutoff' parameter was set on a tenth of the 'exp_cov' value for the final Velvet assembly. Next, a second assembly was produced for each sample using MIRA (21) for correction purposes. The same reference sequence, mapping, and accurate flags were described for Velvet assembly, and lossless digital normalization (-ldn) was activated to reduce the dataset to a memory-manageable size.

Correction, scaffolding and annotation. The viral contigs in both assemblies were selected and ordered by aligning them to the reference sequence using Mauve (22). We also manually merged neighbor contigs in the Velvet assemblies exhibiting a 10 to 90nt long overlap with each other. For each sample, a multiple alignment comprising the GK18 reference sequence, the Velvet assembly, and the MIRA assembly was performed using Kalign2 (23). To generate the final sequences, conserved regions between the Velvet and MIRA assemblies were accepted and the discrepancies were manually corrected by comparing the original read alignments to input sequences. Gap regions were left as strings of NNNs when none of the assemblies could resolve them. We then used Tablet (24) and Jalview (25) for visualization of reads and genome alignment respectively.

After determining the final assembly, each new genome was annotated. Given the close proximity to the KSHV reference sequence, the annotation of each new genome was carried out by transferring the available KHSV annotation in Genbank using RATT (26). However, splicing junctions were unsuccessfully transferred for the K8 and K15 genes. Manual correction was therefore performed for both genes because the K8 exons/introns annotated in the GK18

Genbank entry do not accurately reflect previously reported splicing (27), and the K15 gene features many exons that RATT could not transfer.

Accession numbers. The final consensus sequence of each sample, and it's annotation, can be found at the following accession numbers : KT271453 (ZM004), KT271454 (ZM027), KT271455 (ZM091), KT271456 (ZM095), KT271457 (ZM102), KT271458 (ZM106), KT271459 (ZM108), KT271460 (ZM114), KT271461 (ZM116), KT271462 (ZM117), KT271463 (ZM118), KT271464 (ZM121), KT271465 (ZM123), KT271466 (ZM124), KT271467 (ZM128), KT271468 (ZM130).

Comparative and phylogenetic sequence analysis. The assembled KSHV genomes were aligned using Kalign2. The multiple alignment was then used to generate whole-genome maximum likelihood phylogenetic trees using PhyML (28) with 1000 bootstrap replicates, and the trees were visualized using MEGA6 (29). Genome-wide mutations were visualized using the mVista software (30), with a 100bp scanning window. Multiple amino acid alignments were generated for each KSHV gene and inspected for high level of nonsynonymous mutations. Maximum likelihood phylogenetic trees were generated using PhyML and amino acid highlighter plots were generated to visualize the specific mutations using the Highlighter tool as part of the Los Alamos National Laboratory HIV sequence database (www.hiv.lanl.gov). Reference sequences used for K1 genotyping were as follows: AF133038 (A1), AF130305 (A2), U86667 (A3), AF133039 (A4), AF178823 (A5), AF133040 (B1), AY042947 (B2), AY042941 (B3), DQ309754 (B4), AF133041, (C1), AF133042 (C3), AF133043 (D1), AF133044 (D2), and

AF220292 (E). Reference sequences used for K15 classification were AAD46505.1 (P), AAD45296.1 (M), and a personal communication from Gary Howard (N).

RESULTS

Summary of sequencing data. In the present study, we analyzed the sequence diversity of KSHV whole genomes acquired from KS skin biopsies of sixteen Zambian patients—11 males and 5 females. Clinical data collected for each patient is summarized in Table 1. Total DNA was extracted from the biopsy samples and KSHV burden in tumor tissue ranged from 0.21 to 17.16 copies per cell (Table 2). In order to efficiently sequence the KSHV DNA, which was present as a small proportion of total tumor DNA (0.0006% - 0.05%), we used a custom biotinylated RNA-library (Agilent) to hybridize and selectively enrich the KSHV DNA for sequencing with an Illumina HiSeq.

We first tested the efficiency of our whole protocol, from DNA enrichment to genome sequencing, on three samples (ZM116, ZM117, and ZM118). This process resulted in up to 12,107-fold enrichment of KSHV DNA over cellular DNA with 62% of the total sequence reads mapped to the KSHV reference sequence (GK18). The GK18 sequence was selected because it is currently the most comprehensively annotated KSHV sequence and also served as the reference for generating the RNA bait library employed in the current study. After our initial results, we continued with the remaining 13 samples in a second run of DNA enrichment and deep-sequencing. A total of 528,849,840 paired-end reads, 101bp long, were produced from both HiSeq runs with an average of 49% of the sequence reads mapped to KSHV. Together, we obtained a mean enrichment of KSHV DNA over cellular DNA of 8,437-fold (range 1,851 -

18,235-fold) (Table 2). Thus, we were able to filter the dataset at a high quality threshold and still maintain high depth of coverage (mean: 8,437-fold; range: 786 - 24,740-fold) (Table 2).

The assembly of each viral genome was conducted in a two-step approach: first the sequence reads were assembled using Velvet and the KSHV reference sequence; then this initial assembly was manually corrected with the help of a second assembly generate using MIRA. After manual fusion of contigs exhibiting large overlaps in the initial Velvet assembly, each genome featured an average of 4 contigs. This is consistent with the presence of 3 major repeat regions in KHSV that were hard to resolve using short-read sequencing technology alone. After manual correction with the MIRA assembly, we were able to reduce this value to an average of 2 contigs per genome. All genomes were correctly sized with an average of 137Kb, corresponding to the size of previously published KSHV genomes from Western sources. Apart from the repeat regions, most genomes (12 of 16) had uniform read coverage. However, four genomes (ZM091, ZM095, ZM116, and ZM124) showed a few regions with coverage of up to 3 times the sample average. Nevertheless, these discrepancies did not hinder the assembly process.

KSHV whole-genome variability analysis. The sixteen newly assembled and annotated KSHV genomic sequences, in addition to the six previously sequenced KSHV genomes, were used for multiple alignment and analyzed phylogenetically. Gaps and/or repeated regions of each genome, including the reference sequence, were masked. Figure 1 presents an unrooted maximum likelihood tree depicting the relative phylogenetic distance between samples. Although the overall identity at the nucleotide level is very high among the 22 genomes (Supplemental Figure 1), distinct phylogenetic clusters are evident. All previously published

KSHV genomes from Western countries clustered together, while the isolates from Zambia appear to form two separate clusters and contain much more variability among isolates. Isolate ZM004 diverged substantially from all other sequences and therefore was used to root subsequent whole-genome cladograms. The distinct phylogenetic clusters with higher variability among Zambian isolates can also be seen in the ZM004-rooted cladogram (Figure 2A).

The K1 and K15 genes are known to vary greatly among KSHV isolates and K1 subtypes are associated with specific geographical regions. Therefore, we investigated whether variability in K1 or K15 correlated with the KSHV-whole genome variability we detected. To this end, we performed a multiple alignment of all 22 KSHV genomic sequences without the K1 or K15 genes and generated a ZM004-rooted cladogram. Despite removing the highly variable genes, the topology of the phylogenetic tree remained similar to that of the whole-genome KSHV cladogram (Figures 2A and 2B). The only difference was a slight restructuring among the Western isolates, most likely because the BC-1 isolate contains the K15 M allele. Moreover, when nucleotide phylogenetic trees were generated from the K1 and K15 sequences, the tree topology and sample clustering did not correlate with the whole-genome phylogenetic analysis (Figures 2C and 2D). Together this indicates that the phylogenetic clustering detected at the whole-genome level is a function of variability in the central region of the KSHV genome.

We then sought to investigate whether variability in the central region of the Zambian KSHV isolates could be accurately characterized by individual genes or if consideration of the whole region is required. The distribution of nucleotide variability among the central region of all 22 KSHV genomes compared to GK18 was visualized using mVista (Figure 3). We did not find any

areas of high nucleotide variation, other than the K1 and K15 genes; rather, we detected low-level variation throughout the entire central region when the Zambian isolates were compared to GK18 (Figure 3).

The total number of mutations compared to the GK18 sequence, including deletions, insertions, and substitutions, for each of the sixteen Zambian KSHV genomes is summarized in Table 2. We did not identify any correlations between the Zambian KSHV isolate sequence variation and clinical data. Additionally, we did not detect any intra-subject sequence variation, indicating that KSHV within each KS tumor was clonal.

KSHV coding region mutations. Given that the phylogenetic segregation between Western and Zambian KSHV genomic sequences was due to low level variation across the central region of the genome, we inspected all coding sequences for nonsynonymous mutations compared to the GK18 reference genome. Out of the 84 annotated coding regions, we identified six KSHV genes with high levels of nonsynonymous mutations—four within the central conserved region (K4.2, K8.1, K11/vIRF2, and K12/Kaposin) and two previously known to have high variability (K1 and K15). Among the genes within the central region of the KSHV genome, K4.2 contained the highest level of nonsynonymous mutations compared GK18. Phylogenetic analysis of the K4.2 gene revealed similar clusters for samples ZM004, ZM114, and ZM130 compared to the whole-genome analysis, but not for the remaining samples (Figure 4A). ZM091, ZM095, and ZM118 were very similar to GK18, with only three amino acid substitutions, whereas the remaining K4.2 sequences contained more than 20 substitutions and/or significant truncations at the C-terminal end of the coding region (Figure 4B). Sample ZM124 contained a 25 nucleotide

deletion in the K8.1 coding sequence resulting in a frameshift mutation that produced a stop codon very early in the gene (Figure 5B). Additionally, multiple amino acid substitutions, insertions, and deletions were identified within the K11/vIRF2 and K12/Kaposin genes (Figures 5C and 5D).

Since K1 and K15 have previously been demonstrated to be highly variable and are frequently used for KSHV genotyping, we generated amino acid maximum likelihood phylogenetic trees, including reference sequences, to determine the K1 and K15 genotypes of the sixteen Zambian KSHV isolates (Figure 6). All K1 sequences clustered with genotypes A5 (n = 1) or B (n = 15), consistent with previous K1 genotyping of samples from Zambia (17). Within the B genotype, nine samples clustered with the sub-genotype B1, two with B3, and four with B4. The majority of the Zambian KSHV isolates contained the K15 P allele, however, two isolates (ZM095 and ZM128) contained the rare N allele. This is the first time the K15 N allele has been identified in Zambia.

DISCUSSION

Exploring relationships between KSHV sequence polymorphism and disease pathogenesis requires a more complete perspective of the magnitude and breadth of viral sequence diversity in geographical regions with the highest KSHV disease prevalence. However, little is known regarding the whole-genome diversity of KSHV, as only six complete genomes have been sequenced. Moreover, none of those previously published KSHV genomic sequences derive from sub-Saharan Africa—where prevalence of KSHV and KS is the highest. In the current study, we report the enrichment, sequencing, assembly, and analysis of sixteen unique Zambian

KSHV genomes isolated from KS tumors in adults. This study is the first to utilize SureSelect target capture technology to enrich and sequence KSHV DNA from clinical KS biopsies. Using this approach we were able to efficiently enrich and sequence KSHV from all KS tumor biopsy DNA available, including a sample with KSHV viral burden of only 0.21 copies per cell, to obtain enough depth of coverage for sequence analyses. The data generated represents the largest number of KSHV whole-genomic sequences published to date. Additionally, this study is the first to compare multiple KSHV genomes from a common geographical region that is endemic for KS.

Due to the low ratio of KSHV:human DNA in each tumor preparation, targeted enrichment of KSHV DNA was required prior to Illumina deep-sequencing. The enrichment protocol we employed resulted in 1,851 - 18,235-fold KSHV enrichment. This supported an average coverage depth of 8,437-fold, therefore, allowing high confidence in downstream polymorphism analyses. To further improve the accuracy of our whole-genome assemblies, we used MIRA assembly software to corroborate the Velvet-generated assemblies for each genome. Despite its overall accuracy, Velvet still produced a few clearly misassembled regions or unjustified gaps. Therefore, the two-step assembly enabled us to capitalize on Velvet's accuracy while correcting nearly all misassembled regions and gaps with MIRA.

Given the extremely low error-rate of herpesvirus polymerases (31), it is not surprising that previous comparisons of the six KSHV whole-genomes revealed a high level of sequence conservation (15). The genomic conservation might also be anticipated because the sequences were all generated from US or European patient samples, despite derivation from distinct clinical presentations. Multiple alignment of the sixteen newly sequenced Zambian KSHV genomes also showed high conservation. However, a low level variability across the central conserved region

resulted in distinct phylogenetic clustering between the genomic sequences of Zambian KSHV isolates and those from Western countries.

For EBV, the divergent EBNA-3 alleles correlate with whole-genome clustering and serve as adequate surrogates to distinguish between EBV type-1 and -2 (16, 32), but this has not previously been investigated for KSHV. The genes at either termini of the KSHV genome, K1 and K15, have been previously shown to contain higher levels of polymorphism than the rest of the genome. Our Zambian KSHV sequence data also revealed high levels of sequence diversity at these loci. Nevertheless, when K1 and K15 were excluded from the whole-genome multiple alignment and subsequent phylogenetic analysis the topology of the phylogenetic tree did not change. Moreover, the topology of the KSHV whole-genome phylogenetic tree did not correlate with those of phylogenetic trees generated from K1 or K15 alone, suggesting that these genes are poor surrogates for measuring whole-genome KSHV diversity. Conversely, K4.2 phylogenetic analysis showed partial, but not full, correlation to whole-genomic clustering. This suggests that K4.2 may contribute to viral genomic characterization, but again, this single gene does not adequately characterize the whole-genome diversity we detected. Taken together, our data suggest that the whole KSHV genome should be taken into consideration for accurate viral characterization.

Although single gene polymorphisms did not adequately represent the genome-wide KSHV diversity, we identified several nonsynonymous mutations within KSHV protein coding regions that could potentially affect the viral phenotype. Interestingly, all six genes with significant levels of nonsynonymous mutations were gene products uniquely encoded by KSHV. Five of the

variant genes identified in this study are directly immunomodulatory genes and therefore, may lead to differential effects on the host immune response. Of interest, the coding sequences of K4.2 contained the highest level of nonsynonymous variation. K4.2 interacts with pERP1 to inhibit immunoglobulin secretion and increase calcium influx (33). The amino acid domains of K4.2 that are important for these functions are unknown, hence the newly identified variability may be important for interaction with pERP1 or other cellular functions. This is currently under investigation.

K8.1 protein is not known to be directly immunomodulatory, but this virion membrane-associated glycoprotein does have important effects on the KSHV viral life cycle. K8.1 is utilized for KSHV attachment to target cells and induces VEGF expression upon binding (34, 35). Additionally, while K8.1 is necessary for efficient virus egress from infected cells, a K8.1-null mutant virus can still infect HEK293 cells, indicating that K8.1 function may be dispensable for virus entry *in vitro* (35, 36). The nucleotide sequence from sample ZM124 predicted a truncation in the amino acid sequence before the transmembrane domain of K8.1. If the protein were expressed as predicted, it is not clear what effect a K8.1-deficient virion and/or virion-independent, soluble K8.1 would have in the context of an *in-vivo* infection. K8.1 is highly immunogenic compared to other KSHV proteins (37), and exogenous expression of soluble K8.1 induces an interferon response (38). Together, the ZM124 predicted K8.1 may elicit a modified innate and humoral immune response.

Among the 16 newly sequenced KSHV genomes one isolate, ZM091, contained an A5 K1 gene. The A genotype is primarily found in Europe, while the B genotype is found only in sub-Saharan

Africa; therefore, previous analyses of KSHV molecular evolution have suggested that the A5 genotype arose in sub-Saharan Africa as a result of recombination. While we cannot exclude the possibility of this hypothesis, the sequence data obtained in the present study does not provide evidence of chimeric boundaries near the 5' terminus.

Three primary K15 alleles have been previously identified, P, M, and N. There is little variation within allele groups but extreme divergence across alleles. The P and M alleles, for example, have only 30% identity (8). We identified, for the first time, two KSHV isolates from Zambia that contain the rare K15 N allele. Although the N allele is highly divergent from P and M, the signaling motifs for SH2 and TRAF are conserved in all K15 alleles. Recently, it was shown that the K15-P allele activates the alternative NF-κB signaling pathway by direct recruitment of NF-κB inducing kinase (NIK) to a distinct signal sequence (39). However, this sequence is not conserved in either the M or N alleles. This genetic variation may lead to altered levels of K15-induced NF-κB activation and subsequently functional differences between alleles, thus further functional analyses of these K15 alleles is warranted.

In summary, we successfully enriched KSHV from a background of human DNA from KS biopsies using targeted RNA baits. Analyses of the sequences identified a low level variability across the KSHV central conserved region that resulted in distinct phylogenetic clustering between the genomic sequences of KSHV from Zambia and Western countries. Moreover, four genes within this region had significant levels of polymorphisms but did not adequately characterize the whole-genome diversity we detected. Based on the new sequence data in the present study, it is apparent that whole-genome KSHV diversity is greater than previously

appreciated. Although the observed phylogenetic clustering between Western and Zambian KSHV genomic sequences could represent distinct subtypes, more whole-genome sequences are required from additional regions to infer distinct viral subtypes specific to any geographical region. Continued investigation of the KSHV genetic landscape is necessary in order to effectively understand the role of viral evolution and sequence diversity on KSHV gene functions and pathogenesis.

ACKNOWLEDGMENTS

This work was supported in part by the National Institutes of Health (NIH) (RO1 CA75903, T32 AIO60547, and P30 GM103509 to C.W.) and the Fogarty International Center (D43 TW01492 to C.W.). L.N.O was supported in part by the NIH under a Ruth L. Kirschstein National Research Service Award from the National Institute of Allergy and Infectious Diseases and by a Maude Hammond Fling Fellowship from the University of Nebraska-Lincoln., and C.M. was a Fogarty Fellow.

We thank all patients who participated in this study. We also thank Dr. Gary Hayward for supplying the KSHV K15 type N sequence used for genotyping, and Alok Dhar and James Eudy from the University of Nebraska DNA Sequencing Core for technical assistance. The University of Nebraska DNA Sequencing Core receives partial support from the National Center for Research Resources (1S10RR027754-01, 5P20RR016469, RR018788-08) and the National Institute for General Medical Science (8P20GM103427, GM103471-09).

This publication's contents are the sole responsibility of the authors and do not necessarily represent the official views of the NIH.

REFERENCES

1. **Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, Knowles DM, Moore PS.** 1994. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. Science **266:**1865–1869.
2. **Biggar RJ, Chaturvedi AK, Goedert JJ, Engels EA.** 2007. AIDS-Related Cancer and Severity of Immunosuppression in Persons With AIDS. Journal of the National Cancer Institute **99:**962-972.
3. **Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F.** 2015. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. International Journal of Cancer **136:**E359-E386.
4. **Cesarman E, Chang Y, Moore PS, Said JW, Knowles DM.** 1995. Kaposi's Sarcoma–Associated Herpesvirus-Like DNA Sequences in AIDS-Related Body-Cavity–Based Lymphomas. New England Journal of Medicine **332:**1186-1191.
5. **Soulier J, Grollet L, Oksenhendler E, Cacoub P, Cazals-Hatem D, Babinet P, d'Agay MF, Clauvel JP, Raphael M, Degos L, Sigaux F.** 1995. Kaposi's sarcoma-associated herpesvirus-like DNA sequences in multicentric Castleman's disease [see comments]. Blood **86:**1276-1280.
6. **Uldrick TS, Wang V, O'Mahony D, Aleman K, Wyvill KM, Marshall V, Steinberg SM, Pittaluga S, Maric I, Whitby D, Tosato G, Little RF, Yarchoan R.** 2010. An Interleukin-6-Related Systemic Inflammatory Syndrome in Patients Co-Infected with Kaposi Sarcoma-Associated Herpesvirus and HIV but without Multicentric Castleman Disease. Clinical Infectious Diseases **51:**350-358.
7. **Zong J-C, Ciufo DM, Alcendor DJ, Wan X, Nicholas J, Browning PJ, Rady PL, Tyring SK, Orenstein JM, Rabkin CS, Su I-J, Powell KF, Croxson M, Foreman KE, Nickoloff BJ, Alkan S, Hayward GS.** 1999. High-Level Variability in the ORF-K1 Membrane Protein Gene at the Left End of the Kaposi's Sarcoma-Associated Herpesvirus Genome Defines Four Major Virus Subtypes and Multiple Variants or Clades in Different Human Populations. Journal of Virology **73:**4156-4170.
8. **Poole LJ, Zong J-C, Ciufo DM, Alcendor DJ, Cannon JS, Ambinder R, Orenstein JM, Reitz MS, Hayward GS.** 1999. Comparison of Genetic Variability at Multiple Loci across the Genomes of the Major Subtypes of Kaposi's Sarcoma-Associated Herpesvirus Reveals Evidence for Recombination and for Two Distinct Types of Open Reading Frame K15 Alleles at the Right-Hand End. Journal of Virology **73:**6646-6660.
9. **Hayward GS, Zong J-C.** 2007. Modern evolutionary history of the human KSHV genome, p 1–42, Kaposi Sarcoma Herpesvirus: New Perspectives. Springer.
10. **Rezaee SAR, Cunningham C, Davison AJ, Blackbourn DJ.** 2006. Kaposi's sarcoma-associated herpesvirus immune modulation: an overview. The Journal of General Virology **87:**1781-1804.
11. **Neipel F, Albrecht JC, Fleckenstein B.** 1997. Cell-homologous genes in the Kaposi's sarcoma-associated rhadinovirus human herpesvirus 8: determinants of its pathogenicity? Journal of Virology **71:**4187-4192.
12. **Russo JJ, Bohenzky RA, Chien M-C, Chen J, Yan M, Maddalena D, Parry JP, Peruzzi D, Edelman IS, Chang Y, Moore PS.** 1996. Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8). Proceedings of the National Academy of Sciences **93:**14862-14867.
13. **Brulois KF, Chang H, Lee AS-Y, Ensser A, Wong L-Y, Toth Z, Lee SH, Lee H-R, Myoung J, Ganem D, Oh T-K, Kim JF, Gao S-J, Jung JU.** 2012. Construction and Manipulation of a New Kaposi's Sarcoma-Associated Herpesvirus Bacterial Artificial Chromosome Clone. Journal of Virology **86:**9708-9720.
14. **Yakushko Y, Hackmann C, Günther T, Rückert J, Henke M, Koste L, Alkharsah K, Bohne J, Grundhoff A, Schulz TF, Henke-Gendo C.** 2011. Kaposi's Sarcoma-Associated Herpesvirus Bacterial Artificial Chromosome Contains a Duplication of a Long Unique-Region Fragment within the Terminal Repeat Region. Journal of Virology **85:**4612-4617.

15. **Tamburro KM, Yang D, Poisson J, Fedoriw Y, Roy D, Lucas A, Sin S-H, Malouf N, Moylan V, Damania B, Moll S, van der Horst C, Dittmer DP.** 2012. Vironome of Kaposi sarcoma associated herpesvirus-inflammatory cytokine syndrome in an AIDS patient reveals co-infection of human herpesvirus 8 and human herpesvirus 6A. Virology **433:**220-225.
16. **Kwok H, Wu CW, Palser AL, Kellam P, Sham PC, Kwong DLW, Chiang AKS.** 2014. Genomic Diversity of Epstein-Barr Virus Genomes Isolated from Primary Nasopharyngeal Carcinoma Biopsy Samples. Journal of Virology **88:**10662-10672.
17. **Olp LN, Shea DM, White MK, Gondwe C, Kankasa C, Wood C.** 2013. Early childhood infection of Kaposi's sarcoma-associated herpesvirus in Zambian households: A molecular analysis. International Journal of Cancer **132:**1182-1190.
18. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:**2114-2120.
19. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods **9:**357-359.
20. **Zerbino DR, Birney E.** 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research **18:**821-829.
21. **Chevreux B, Wetter T, Suhai S.** 1999. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information, p 45 - 56. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB).
22. **Darling AE, Mau B, Perna NT.** 2010. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLoS ONE **5:**e11147.
23. **Lassmann T, Frings O, Sonnhammer ELL.** 2009. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Research **37:**858-865.
24. **Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D.** 2013. Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics **14:**193-202.
25. **Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ.** 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics **25:**1189-1191.
26. **Otto TD, Dillon GP, Degrave WS, Berriman M.** 2011. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Research **39:**e57-e57.
27. **Lin S-F, Robinson DR, Miller G, Kung H-J.** 1999. Kaposi's Sarcoma-Associated Herpesvirus Encodes a bZIP Protein with Homology to BZLF1 of Epstein-Barr Virus. Journal of Virology **73:**1909-1917.
28. **Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology **59:**307-321.
29. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S.** 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution **30:**2725-2729.
30. **Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I.** 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Research **32:**W273-W279.
31. **Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R.** 2010. Viral Mutation Rates. Journal of Virology **84:**9733-9748.
32. **Dolan A, Addison C, Gatherer D, Davison AJ, McGeoch DJ.** 2006. The genome of Epstein–Barr virus type 2 strain AG876. Virology **350:**164-170.
33. **Wong L-Y, Brulois K, Toth Z, Inn K-S, Lee S-H, O'Brien K, Lee H, Gao S-J, Cesarman E, Ensser A, Jung JU.** 2013. The Product of Kaposi's Sarcoma-Associated Herpesvirus Immediate Early Gene

K4.2 Regulates Immunoglobulin Secretion and Calcium Homeostasis by Interacting with and Inhibiting pERP1. Journal of Virology **87:**12069-12079.

34. **Birkmann A, Mahr K, Ensser A, Yağuboğlu S, Titgemeyer F, Fleckenstein B, Neipel F.** 2001. Cell Surface Heparan Sulfate Is a Receptor for Human Herpesvirus 8 and Interacts with Envelope Glycoprotein K8.1. Journal of Virology **75:**11583-11593.
35. **Subramanian R, Sehgal I, D'Auvergne O, Kousoulas KG.** 2010. Kaposi's Sarcoma-Associated Herpesvirus Glycoproteins B and K8.1 Regulate Virion Egress and Synthesis of Vascular Endothelial Growth Factor and Viral Interleukin-6 in BCBL-1 Cells. Journal of Virology **84:**1704-1714.
36. **Luna RE, Zhou F, Baghian A, Chouljenko V, Forghani B, Gao S-J, Kousoulas KG.** 2004. Kaposi's Sarcoma-Associated Herpesvirus Glycoprotein K8.1 Is Dispensable for Virus Entry. Journal of Virology **78:**6389-6398.
37. **Labo N, Miley W, Marshall V, Gillette W, Esposito D, Bess M, Turano A, Uldrick T, Polizzotto MN, Wyvill KM, Bagni R, Yarchoan R, Whitby D.** 2014. Heterogeneity and Breadth of Host Antibody Response to KSHV Infection Demonstrated by Systematic Analysis of the KSHV Proteome. PLoS Pathog **10:**e1004046.
38. **Perry ST, Compton T.** 2006. Kaposi's Sarcoma-Associated Herpesvirus Virions Inhibit Interferon Responses Induced by Envelope Glycoprotein gpK8.1. Journal of Virology **80:**11105-11114.
39. **Hävemeier A, Gramolelli S, Pietrek M, Jochmann R, Stürzl M, Schulz TF.** 2014. Activation of NF-κB by the Kaposi's sarcoma herpesvirus K15 protein involves recruitment of the NF-κB-inducing kinase, IκB kinases and phosphorylation of p65. Journal of Virology**:**JVI.01766-01714.

FIGURE LEGENDS

Figure 1. Unrooted nucleotide maximum likelihood phylogenetic tree of six previously published KSHV whole-genome sequences and 16 new KSHV whole-genome sequences from Zambian KS biopsies. Phylogenetic tree was generated using PhyML with 1000 bootstrap replicates and visualized using MEGA6.

Figure 2. Maximum likelihood phylogenetic tree analysis of 22 KSHV whole genomes and nucleotide sequence of K1 and K15 genes. All phylogenetic trees were generated using PhyML with 1000 bootstrap replicates and visualized using MEGA6. (A) KSHV whole genome cladogram rooted on sample ZM004. (B) ZM004-rooted cladogram of KSHV whole genome with K1 and K15 gene sequences removed. (C and D) Midpoint-rooted cladogram of KSHV K1 (C) and K15 (D) sequences from 22 KSHV whole genome sequences.

Figure 3. Distribution of nucleotide variation within the central region of the KSHV genome as compared to the reference sequence GK18. The figure was generated using mVista software with a 100bp scanning window. The curve for each sequence represents up to 10% nucleotide variation within that window. Topological phylogenetic tree of sequences was generated using PhyML and visualized using MEGA6. Of note, multiple regions within the DG-1 sequence appear to have significant nucleotide diversity compared to GK18; however, these regions represent gaps in the published sequence due to low coverage that could not be masked.

Figure 4. Amino acid polymorphisms within the K4.2 gene. (A) Maximum likelihood phylogenetic tree of K4.2 amino acid sequence generated using PhyML with 1000 bootstrap

replicates and visualized using MEGA6. (B) Amino acid highlighter plot of K4.2 generated using the Highlighter tool as part of the Los Alamos National Laboratory HIV sequence database.

Figure 5. Amino acid polymorphisms of three other KSHV central region genes. (A) Amino acid highlighter plot coloring scheme. Amino acid highlighter plots were generated for KSHV genes K8.1 (B), vIRF-2 (C), and K12 (D) using the Highlighter tool as part of the Los Alamos National Laboratory HIV sequence database.

Figure 6. KSHV K1 and K15 subtyping. Phylogenetic analyses using amino acid sequences for K1 (A) and K15 (B) from the 22 KSHV genomic sequences and prototypic reference sequences. Maximum likelihood phylogenetic trees were generated using PhyML with 1000 bootstrap replicates and visualized using MEGA6.

Journal of Virology

Table 1. Clinical information for 16 KS patients.

Patient ID	Gender	Age	HIV Status	ART Status
ZM004	F	35	+	+
ZM027	F	15	+	-
ZM091	M	30	+	+
ZM095	M	41	+	+
ZM102	F	45	+	N/A
ZM106	M	36	+	+
ZM108	M	33	+	+
ZM114	F	29	+	N/A
ZM116	F	42	N/A	N/A
ZM117	M	34	+	N/A
ZM118	M	29	N/A	N/A
ZM121	M	37	+	+
ZM123	M	30	+	+
ZM124	M	25	N/A	N/A
ZM128	M	32	+	-
ZM130	M	33	+	-

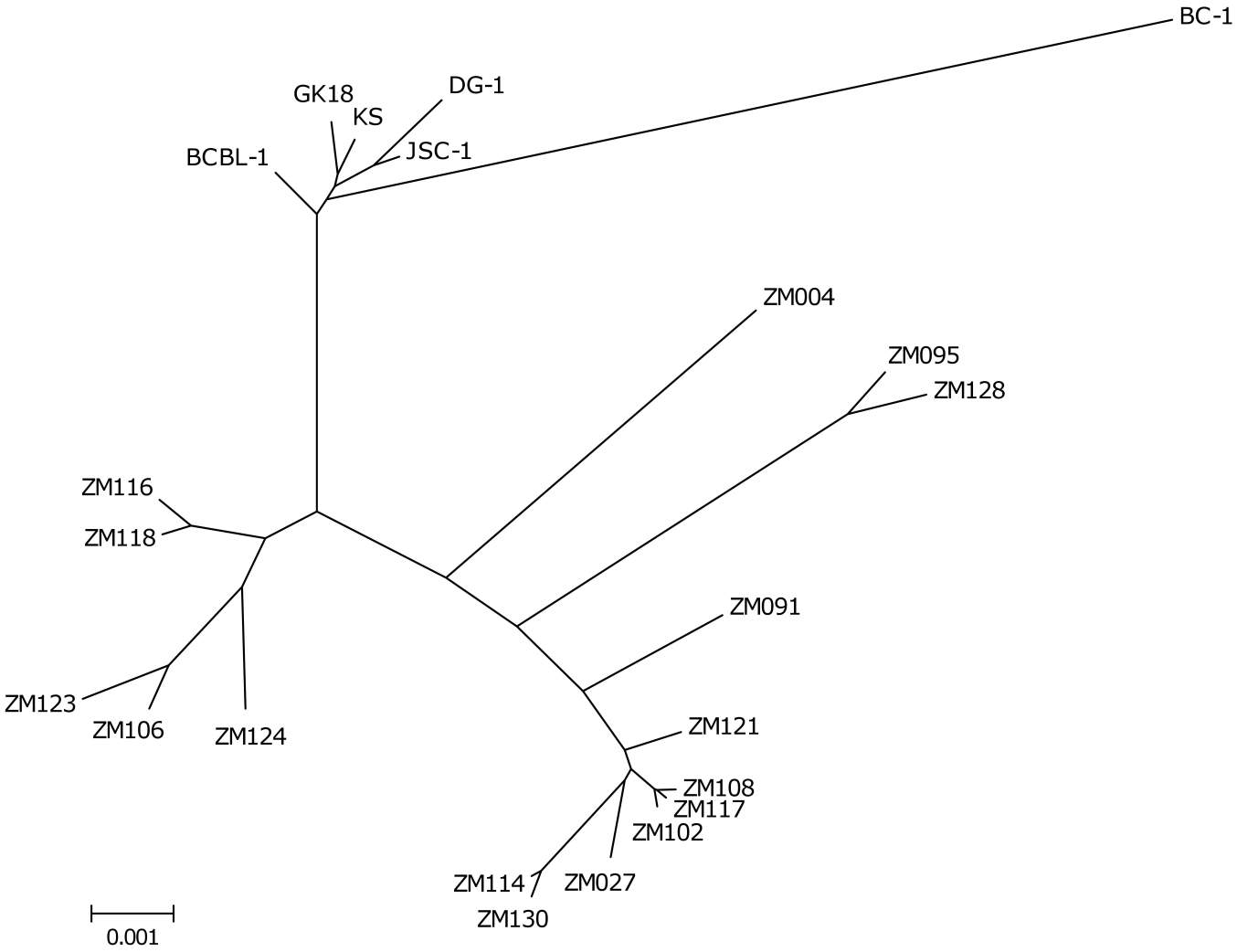
N/A, Not available.

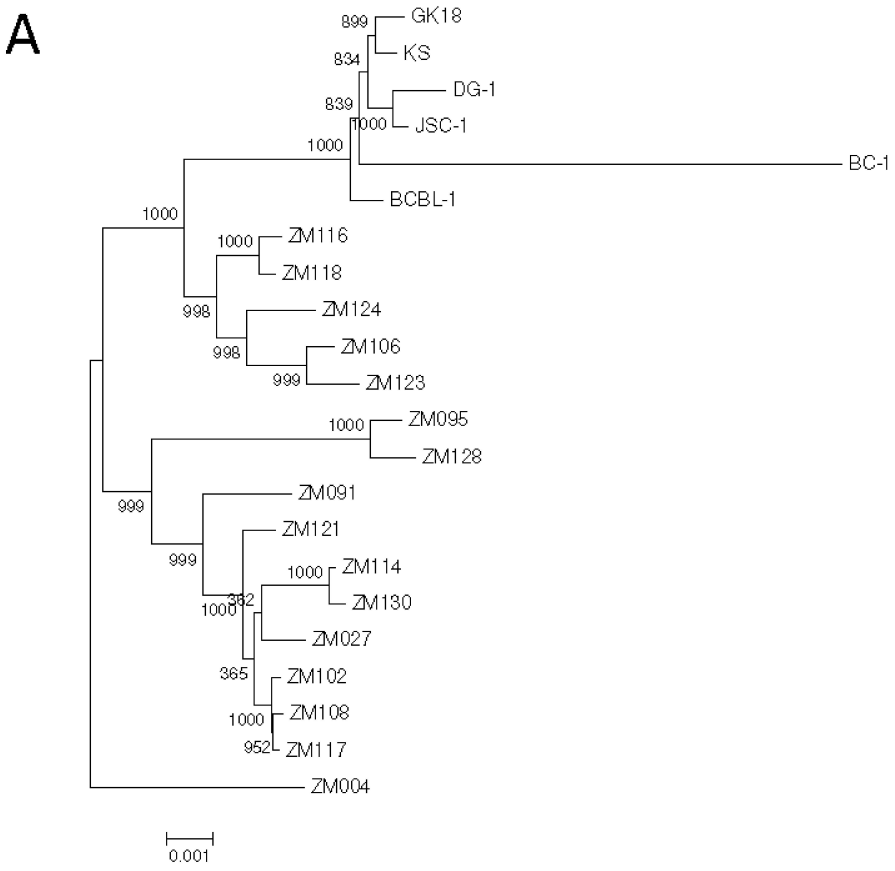
Table 2. Summary of data from sequencing analysis.

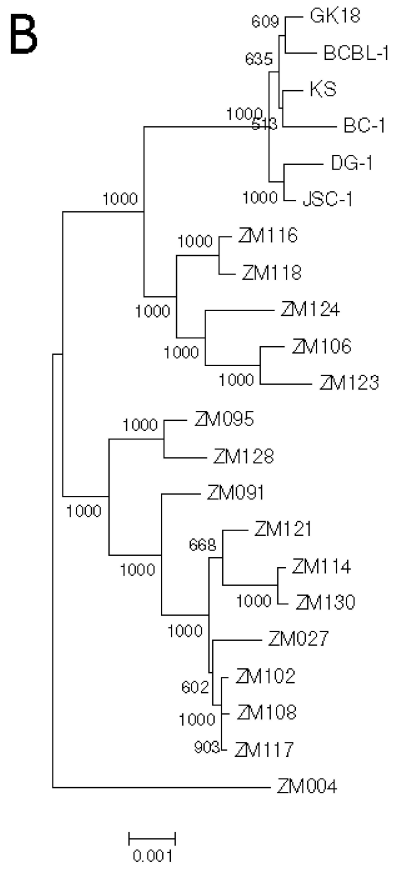
Patient ID	KHSV Genome Size (bp)	Depth of Coverage	Total mutations	Insertions	Deletions	Substitutions	KSHV copies per cell in biopsy DNA	KSHV:Human DNA Before Enrichment	KSHV:Human DNA After Enrichment	Fold Increase
ZM004	136,691	4,507	1344	63	58	1223	1.43	0.0038%	38.43%	10,184
ZM027	137,054	9,160	1289	63	61	1165	5.51	0.0146%	66.50%	4,546
ZM091	136,919	4,448	1063	53	61	949	0.76	0.0020%	37.03%	18,235
ZM095	137,610	9,886	1714	104	104	1506	17.16	0.0456%	84.36%	1,851
ZM102	136,629	8,228	1214	53	68	1093	3.84	0.0102%	59.51%	5,830
ZM106	137,026	7,570	1032	67	66	899	9.68	0.0257%	67.16%	2,611
ZM108	136,568	7,758	1196	62	63	1071	2.72	0.0072%	53.72%	7,438
ZM114	137,071	3,861	1237	61	42	1134	1.4	0.0037%	40.40%	10,841
ZM116	136,969	24,740	862	39	39	784	1.95	0.0052%	62.75%	12,107
ZM117	137,143	15,517	1183	55	61	1067	2.07	0.0055%	57.85%	10,541
ZM118	137,429	19,961	859	58	18	783	ND	ND	45.21%	ND
ZM121	137,279	1,023	1250	119	63	1068	ND	ND	8.52%	ND
ZM123	136,262	786	1404	423	67	914	0.21	0.0006%	7.08%	12,768
ZM124	137,457	6,209	911	43	58	810	ND	ND	73.84%	ND
ZM128	137,272	6,740	1858	200	89	1569	2.75	0.0073%	51.64%	7,076
ZM130	136,908	4,603	1259	54	72	1133	1.14	0.0030%	32.74%	10,822

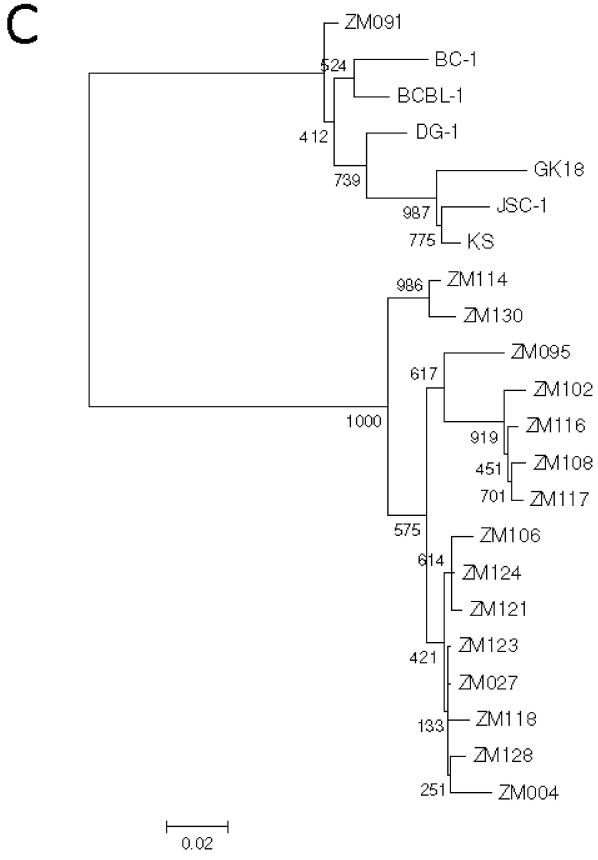
ND, Not determined.

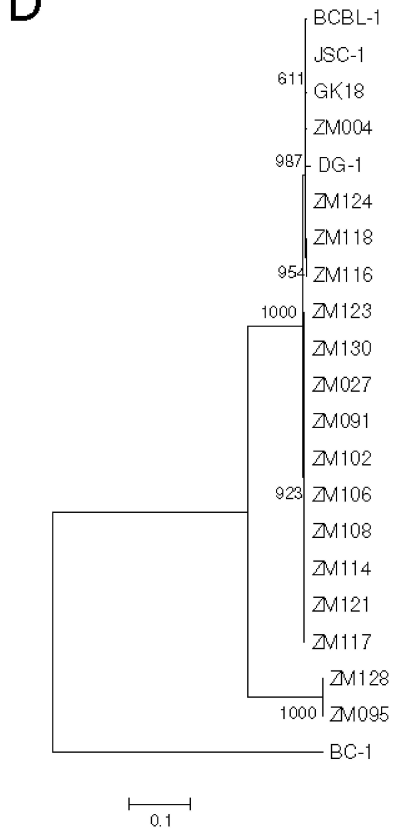
Journal of Virology

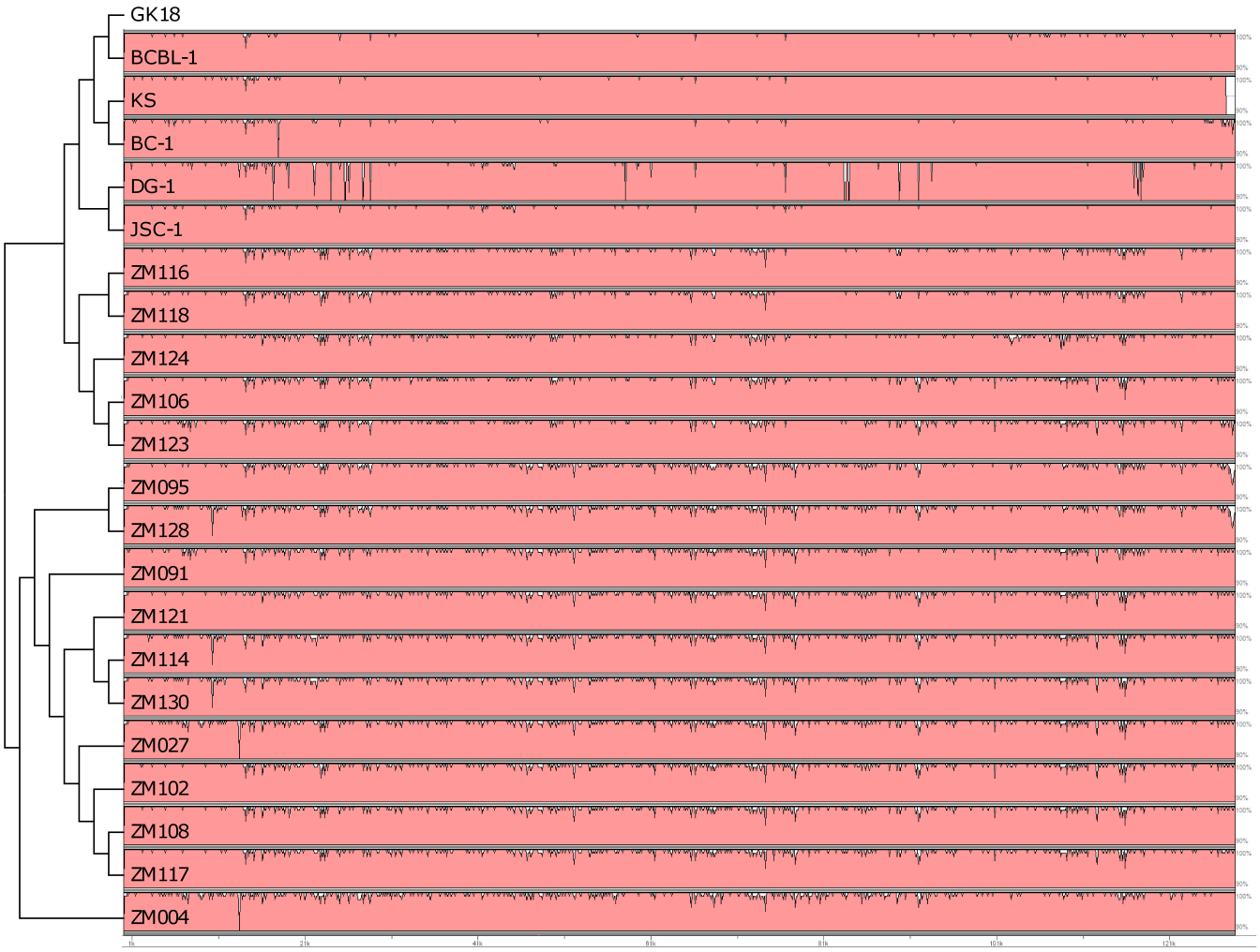


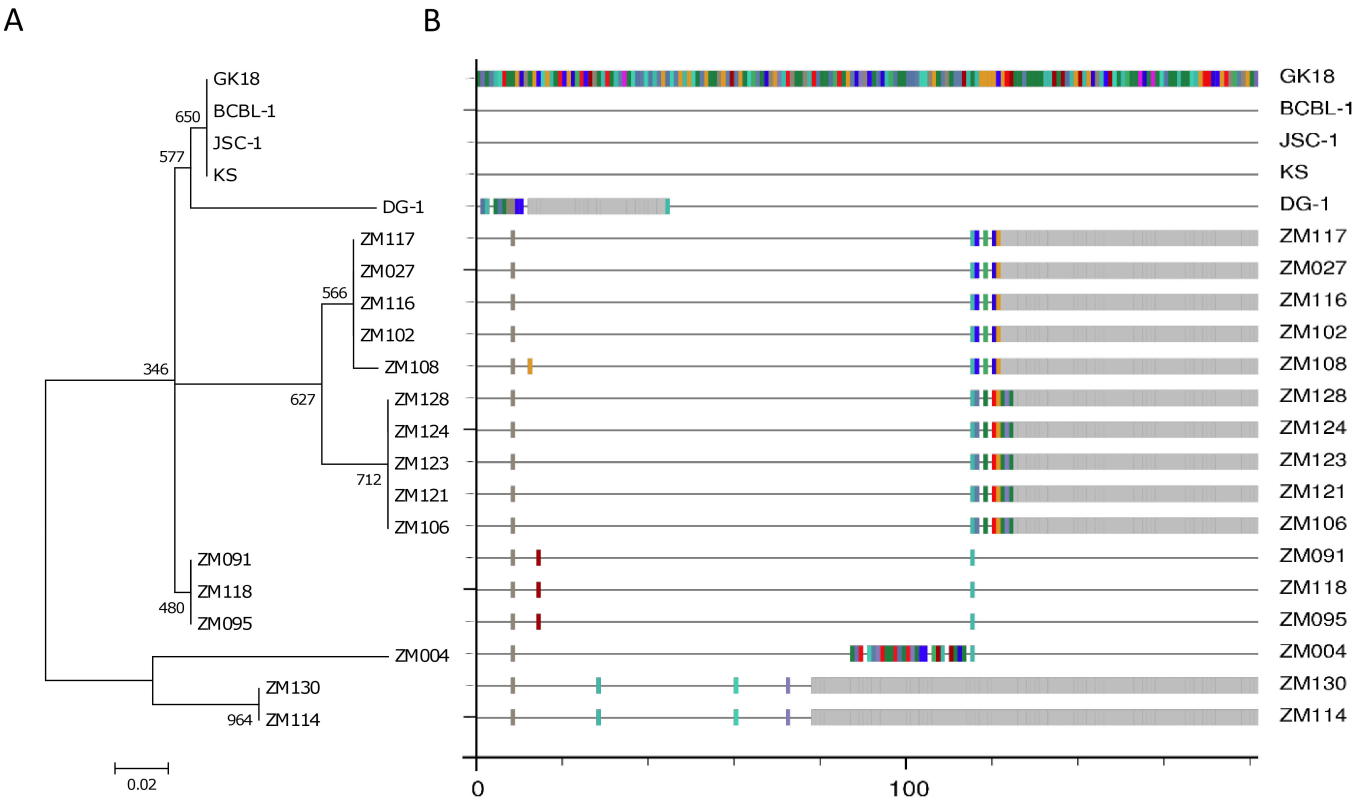












A